IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 April 2023, accepted 10 May 2023, date of publication 16 May 2023, date of current version 23 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3276761

RESEARCH ARTICLE

Synthesis of Tax Return Datasets for Development of Tax Evasion Detection

NARONGCHAI VISITPANYA AND TAWEESAK SAMANCHUEN
Technology of Information System Management Division, Faculty of Engineering, Mahidol University, Phuttamonthon, Nakhon Pathom 73170, Thailand
Corresponding author: Taweesak Samanchuen (taweesak.sam@mahidol.ac.th)

ABSTRACT Datasets are an essential part of data science processes. However, retrieving a dataset, especially a tax return dataset, is challenging as privacy becomes more evident in our daily lives. Thus, data synthesis is an approach selected for our work by utilizing publicly available data and augmenting it using Generative Adversarial Network (GAN) and Synthetic Minority Oversampling TEchnique (SMOTE). The evaluation is performed using a Correlation Matrix, Principal Component Analysis (PCA), and Quality Score. In addition, fundamental machine learning models are utilized to detect tax evasion based on a literature review. The data are gathered from the financial statements of companies registered within the Stock Exchange of Thailand (SET). Our results indicate that synthetic datasets with 0.87 average Quality Score can train models that yield approximately 0.95 Accuracy and 0.91 F1-Score. Additionally, by increasing more instances, the effect of class imbalance and high variance can be mitigated. The expected benefits include the use of open data for analysis and application of synthetic datasets. Forthcoming research could consider the statistical behavior of different business sectors, multiclass labeling for advanced recommendations, and implementation of unsupervised models.

INDEX TERMS Synthetic dataset, tax evasion, financial statement, GAN, SMOTE.

I. INTRODUCTION

Data has become a fundamental asset for every business unit, including a government. With modern technologies, several variation of data products have been created to enhance the competitive advantage of the company. Because of its benefits, the demand for data increases while data usage needs to be controlled as the dimension of privacy rights. Several laws and regulations have been implemented to ensure privacy [1].

Tax return data are crucial data that can show the discipline of a person or a company on tax payments. The government needs to make sure that the performance of tax collection is as high as possible. With modern machine learning techniques, tax payment behavior can be determined, which has become an exciting research topic. However, tax return data is protected data. Requesting this data from the data owner, a government office, becomes a complicated process, even for educational purposes. Based on this issue, several works tried to generate data by collecting data, extracting knowledge from publicly available sources, and interviewing experts [2], [3]. In this work, the financial statements of companies listed in the Stock Exchange of Thailand (SET) are used [4].

Tax is an obligation of citizens who earn a living in a country to make a payment to the government according to the law to support the state and its affairs [5]. Suppose an individual or corporation attempts to avoid paying tax by underreporting income or overstating deductions and exemptions. This behavior is called "Tax Evasion" and is considered an illegal practice [6], [7], [8]. Tax Evasion can also be classified as a "shadow economy" because governments lose revenue from implementing economic and social policies. They also need to spend this revenue to detect fraud [9], [10], [11], [12]. In addition, such behavior is unfair to compliant taxpayers [6]. Therefore, fighting fraud is inevitable for a government. From a statistical perspective, in 2021, the world lost over 480 billion USD owing to tax abuse committed by multinational corporations and individuals. This loss can provide complete vaccination for the world population over three times [13].

As tax declarations and related documents obtained from organizations are enormous, it is crucial to use technology to conduct audits. Moreover, tax evasion techniques change

The associate editor coordinating the review of this manuscript and approving it for publication was Wai-Keung Fung.

over time [14], the honesty of tax officials is not stable, or a bribe is still happening. Thus, digitalization can strengthen tax auditing tasks and mitigate tax evasion [8].

Business Intelligence (BI) can help tax personnel screen financial statements and select suspicious ones for further audit. Additionally, the use of machine learning can help officials address these problems. The model can reflect insights more profoundly than rigid risk criteria can. Machine learning can be categorized into two general types: supervised and unsupervised learning. The main difference is the necessity of labels, which is compulsory for training a model. Supervised learning requires these data to develop a model, whereas unsupervised learning does not [15]. Integrating digitalization and machine learning can support officials in transmitting tax data smoothly between divisions and in exploring new and evolving tax evasion cases [14]. Moreover, there are other categorizations of machine learning techniques, such as semi-supervised and reinforcement learning [15], [16], [17]. These techniques can also be used in this application.

Data are a major element to train machine learning. Importantly, they need to be handled properly as some of them are sensitive such as personal data, financial data, or health-related data. Thus, many laws and regulations that may be written either within organizations or by external organizations, such as government sectors, have been introduced to ensure appropriate utilization of such data. Therefore, it is necessary to consider regulations before using or sharing confidential data. In terms of internal regulations, each company has a policy to protect the privacy of customer data. For external regulations, data processors must be aware of such rules before processing customer data. Several laws and regulations regarding privacy are, for example, the GDPR in EU countries [18], several in the US, e.g., HIPPA, FCRA, FERPA, or GLBA [19], and PDPA in Thailand [20]. This means that obtaining the actual data is cumbersome. If obtaining them is possible, it is still mandatory to consider restrictions from authorities such as data subjects, data owners, or data controllers. Alternatively, financial statements from other stock markets such as NASDAQ [21], Frankfurt Stock Markets [22], and SET [4] are publicly available. Knowledge can be gathered for further work in this area. Additionally, financial statements from SET underwent an auditing process before being publicly available. Thus, it could be implied that knowledge from such data represents the natural behavior of financial data. Some advantages of using this concept are the reduction of disclosure risk, derivation of data to tailor research questions, and making adjustments to data behavior for machine learning model building [23], [24], [25].

In short, the secretiveness of data, tax evasion behavior, and loss of integration between departments are the main pain points that lead to an exaggeration of the problem [1], [8], [14]. As previously mentioned, implementing digitalization can mitigate the severity of this problem. Moreover, privacy issues play an essential role in this work because raw data cannot be easily obtained owing to restrictions. In terms of the overall tax loss due to global tax abuse, the world has lost 483 billion USD by 2021 [13]. Therefore, it is crucial to conduct a study in this area to provide guidelines for gathering data to analyze and alleviate the impact of tax evasion on society. Our motivation is to create a process for gathering data from public sources and combining them into tax return datasets to construct a machine learning model based on such data. This can support decision making in areas where it is difficult to obtain actual data. Our contributions are as follows:

1) To synthesize tax return datasets by studying financial statements and obtaining them from public sources.
2) To increase the number of instances through Generative Adversarial Network (GAN) and Synthetic Minority Oversampling TEchnique (SMOTE) for further analysis.
3) To demonstrate the usability of synthetic datasets by using them to train models.

Our scope of work is as follows:

1) This work considers tax evasion cases based on Corporate Income Tax (CIT) and Value-Added Tax (VAT) using the Thai tax system as a case study.
2) The dataset is created by gathering knowledge from a literature review and the SET website [4].
3) Suspicious cases are based on the risk assessment criteria selected from the literature review.

The article begins with a literature review in Section II. This Section explains a definition of tax-related terms, financial statements, data synthesis approaches, and supervised learning. Then, in Section III, our processes, including acquiring data, augmenting data, and applying machine learning models, are introduced. After that, in Section IV results are evaluated and discussed. Finally, in Section V, our work is concluded to recap important ideas and discuss future work.

II. LITERATURE REVIEW

This section introduces related knowledge for our work. This covers tax knowledge, audit processes, data sources, data synthesis, supervised learning, and related work.

A. TAX KNOWLEDGE

First of all, we introduce general tax knowledge. In this Section, we begin with a definition of tax-related terms and discuss tax evasion behavior.

1) TAX DEFINITION

According to the Ministry of Finance [5], tax is an obligation of citizens to pay to the government to support the state and its affairs. In other words, it is a primary source of fiscal revenue. It is one of the resources the government uses to drive the country forward.

2) TAX EVASION

Tax evasion is an illegal practice, in which one attempts to avoid paying taxes or to reduce the tax base to pay less. According to the Organisation for Economic Co-operation and Development (OECD), there are two obvious forms of tax evasion: electronic suppression, in which under-reporting is considered, and false invoicing, in which over-deduction is considered [6].

There are also behavior that are considered tax evasion [26]. First, a company that is established in a duty-free territory indicate a risk of tax evasion. Because the sales and financial information are not exchanged with external parties. Second, the creation of non-existing materials, which include documents, transactions, and tax reduction documents, is considered as tax evasion behavior and the so-called false invoicing [6]. Third, under-reporting income is also classified as tax evasion behavior [6]. Fourth, doing business with an affiliated party at a high or low price to reduce taxes is considered tax evasion [27], [28]. In addition, selling with an affiliated party with a non-existing bill or a specific bill for tax reduction is classified as tax evasion. Finally, money laundering is an alternative form of evading tax through a shell company, a type of actual company, to launder money. This means hiding the actual income gained from committing a tax crime or using illegally earned money for a typical transaction [29].

In doing business, the institutional theory is an important aspect of tax, which consists of informal and formal institutions [7]. An informal institution is based on ethical behavior and the unwritten rules that society accepts. Simultaneously, formal institutions create trust by the auditing processes, laws, and written regulations that people must obey. An informal mechanism is important for a company's image. Formal institutions must consider their audit strength, laws, and regulations. After a company undergoes an audit, it must consider the consequences of the action according to the law and regulations if there are unlawful actions.

In contrast, if any person or organization performs any activity that reduces the tax liability within the legal framework, such as donations to educational or health organizations, purchasing life or accident insurance, or participating in other government measures, this is called ''Tax Avoidance'' [30], [31]. In summary, the key difference between tax evasion and tax avoidance is that the former is illegal whereas the latter complies with law.

B. AUDIT PROCESS

Previously, we present the definition and behavior of tax evasion. In this Section, we explain more information about auditing processes as tax evasion behavior surreptitiously exists.

1) GENERAL AUDIT PROCESS

Traditionally, fraud detection is based on rule-based expert systems. In this approach, rule and static criteria filter out suspicious behaviors, where a risk index or score is given to represent the chance and severity of fraud behavior in each case [32], such as the observation of the liquidity of cash flow, profit gain in any period, or amount of tax exemption.

Common methods for tax inspection include manual, digital, and whistleblower-based selection [16], [33]. First, the rules and criteria for manual selection are listed. Officials audit them either manually or on a case-by-case basis. However, this method is time-consuming and expensive. In addition, many fraudsters have developed trickier strategies to avoid inspections by tax authorities [32], [34]. Second, many countries apply digital selection, such as data mining, as the primary approach to performing an audit because it is time efficient and cost saving. According to OECD report, many countries have integrated electronic devices into the cash registers of enterprises to collect all data for an audit. For example, Austria implemented a secure signature in a cash register and Quebec, a province in Canada, developed a Sales Recording Module (SRM) [6]. Finally, whistleblower-based selection is a method based on notifications from informants to inform tax authorities about fraudulent behavior. For example, according to the Internal Revenue Service of the United States [35], up to 15% of the collected taxes and penalties can be granted to informants. In Thailand, the Revenue Department has a channel for the Whistle Blower through an online portal [36].

2) RISK ASSESSMENT

As previously mentioned, the Thai tax system is selected as our case study. In this subsection, we briefly introduce Thailand's auditing process. The audit process for tax case selection, which the Revenue Department uses, is called ''Risk Assessment.'' This risk assessment criterion goes through research, and primarily indicates the probability of suspicious cases. Therefore, further analysis is required before selecting tax evasion cases.

Table 1 presents examples of the risk assessment criteria. Not all cases corresponding to the risk assessment criteria are processed through in-depth audits due to the limitation of the resources. To select a case for further auditing, the weighted score in Table 1 is assigned to an instance that corresponds to a criterion. An instance can be subjected to more than one criterion. Other factors include the type of business that can be identified using the International Standard Industrial Classification for all economic activities (ISIC), size of the business, and income [37]. After all instances received a weighted score, they are ranked in order based on their overall scores. Instances from the top ranks are prioritized.

C. DATA SOURCES

As many laws and regulations regarding financial data are enforced, nowadays, the usage of these data is restricted. In this Section, we explain some possibilities to find ones that are available publicly.

1) OPEN DATA

To use open data, one can access several websites such as Kaggle, the UCI Repository, and Google Data Search. There are only a few examples and additional sources provide

TABLE 1. Risk assessment criteria of tax evasion; retrieved from Table 2.1 in [37].

No.	Risk Criteria Description	Weighted Score
1	Purchase volume more than sale volume	3
2	Individual P/T more than regional P/T	2
3	Sale increase while tax decrease	2
4	Claim credit for six months in a row	4
5	Income increase while tax decrease	1
6	Gross loss	5
7	Inventories more than direct income	4
8	Products at the end of the period more than buying products	4
9	No fixed asset and not rent	1

publicly available data. For example, some organizations may publish annual reports for investors or interested people.

When using open-source data, verifying the reliability of the data or data quality is compulsory to obtain reliable results from the data; for example, consideration of data type, contact person, data source, data format, and year of publication. Under our circumstances, the financial statements of enterprises listed in SET [4] are analyzed, as mentioned in our scope of work. In general, international stock markets, such as the NASDAQ [21] or Frankfurt stock markets [22], also provide the financial statements of companies listed in such stock markets. These can also be used to construct a dataset to tailor research questions.

2) FINANCIAL STATEMENT

To extract knowledge from financial statements available from companies listed in SET [4], we should consider the reliability of the data. Data from these sources are audited before being publicly available. To understand the attributes required for the analysis, this section discusses several definitions of the attributes in a financial statement.

According to Table 2, the following are the definitions of attributes [38].

1) Revenue
   a) Sale, R_s, is a sale of products or services. It is also called "Direct Income."
   b) Return, R_r, is an attribute that must be reduced from sales revenue. If there is a high return rate, it is implied that the quality of products may be an issue.
   c) Total Revenue, R_t, is defined as:

$$R_t = R_s - R_r. \tag{1}$$

2) Expense
   a) Inventory is the value of the remaining goods intended to be sold to the consumers. According to Table 2, there are two types of inventories: Beginning Inventory, I_b, and Ending Inventory, I_e. These are used to indicate the goods that remain at the beginning and end of the accounting year.
   b) Goods for Sale, I_s, is the number of goods in a warehouse used to sell between cycles, which is defined as

$$I_s = I_b + E_p, \tag{2}$$

   where E_p is a value of goods purchased during the cycle.
   c) Cost of Goods Sold, E_c, is the cost of good sold in the cycle and is used for calculating "Gross Profit." E_c is defined as

$$E_c = I_s - I_e. \tag{3}$$

3) Gross Profit (P_g) is the initial profit calculated using R_t and E_c. If the attribute "Gross Profit" is less than zero, then it represents loss and is referred to as "Gross Loss." P_g is defined as

$$P_g = R_t - E_c. \tag{4}$$

4) Operating Expense
   a) Selling Expense, E_s, is the expense of selling goods such as advertisements, sales commissions, or trial goods. This type of expense should be excluded when calculating Gross Profit (4).
   b) Administrative Expense, E_a, is an operating expense, such as the salary of staff, excluding sales staff and their commission, facilities expenses, or utility expenses. Similarly, to calculate the Gross Profit, this type of expense should be excluded.
   c) Total Operating Expense, E_t, is defined as

$$E_t = E_s + E_a. \tag{5}$$

   d) Operating Profit, P_o, is defined as

$$P_o = P_g - E_t. \tag{6}$$

5) Net Profit or Loss
   a) Interest, E_f, is the financial expense that companies must pay to creditors who provide short-term or long-term loans.
   b) Profit before taxation, P_{bt}, is the profit that companies gain after deducting their revenue from all expenses. P_{bt} is defined as

$$P_{bt} = P_o - E_f. \tag{7}$$

   c) Income Tax, T_i, is calculated based on the type of company. The amount of tax paid depends on the type of business and tax rate. For example, if companies registered within SET [4] are considered, the rate is 20% [39].
   d) Net Profit, P_n, is the profit companies gain from doing business after deducting all expenses and income tax. P_n is defined as

$$P_n = P_{bt} - T_i. \tag{8}$$

This subsection presents a general overview of financial statements. Financial statements from different sources may

TABLE 2. An example of financial statement.

	Financial statement **From the year ended 31 December 2014** (in Baht)
Revenue	
Sales	1,750,000
Return and Discount	50, 000
Total Revenue	1, 700, 000
Expense	
Beginning Inventory (01 JAN 2014)	150,000
Purchase	1, 050, 000
Goods for Sales	1,200,000
Ending Inventory (31 DEC 2014)	200, 000
Cost of Goods Sold	1, 000, 000
Gross Profit (Gross loss)	700, 000
Operating Expense	
Selling Expense	150,000
Administrative Expense	100, 000
Total Operating Expense	250, 000
Operating Profit	450, 000
Net Profit (Loss)	
Interest Expense	20,000
Profit before Tax	430,000
Income Tax	86, 000
Net Profit (Loss)	344, 000

differ from the examples mentioned in this subsection; however their attributes should be similar. Otherwise, financial equations can be used to solve unknown attributes. Next, we discuss data synthesis where data can be synthesized for more instances.

D. DATA SYNTHESIS

Occasionally, it is possible to obtain data from public sources. However, if we can only collect a limited amount of data, we can synthesize them to obtain more instances. In this section, we discuss GAN and SMOTE techniques that assist us by generating more data.

1) GENERATIVE ADVERSARIAL NETWORK

The previous subsection presented a method for retrieving data by accessing public data. However, if a data source can be accessed with a limited amount of data, the data can be augmented or synthesized to obtain sufficient instances. A well-known method for achieving this is to use GAN to capture the original statistical characteristics of a real dataset and synthesize a new dataset that represents such characteristics.

GAN is an artificial intelligence (AI) algorithm for solving problems related to generative modeling [40]. It comprises two neural networks: a "generator" and "discriminator." New samples are generated by capturing the main distribution of real samples using the generator. Subsequently, the generated new samples are differentiated from the real samples, which is performed by the discriminator [41], [42], [43]. Note that only the discriminator can access both the original and machine-generated data, whereas the generator can only access the characteristics of real instances, not a dataset, and generate data based on them [43].

One type of dataset that is widely used today is tabular data such as demographic data, financial data, and medical records. As privacy becomes increasingly important, careful consideration is required when using data for analysis. Consequently, the data and their quantities are limited to third parties. This problem can be lessened by applying GAN in which privacy is still respected while instances of data for analysis are sufficient. There are two types of GAN for tabular data, which are discussed in this subsection. These are Tabular GAN (TGAN) [43] and Conditional Tabular GAN (CTGAN) [45].

TGAN comprises two neural networks: a generator that generates synthetic data and a discriminator that classifies the data synthesized by the generator from the original dataset. In this model, the generator is based on a long short-term memory (LSTM) network, whereas a multilayer perceptron (MLP) is an algorithm for the discriminator. The performance of TGAN can be evaluated by considering machine learning performance and a correlation matrix between an actual and synthetic dataset [44].

TGAN provides a model to synthesize tabular data, but it does not emphasize the class distribution in a dataset. CTGAN is developed to consider a predetermined class. To solve this problem, CTGAN uses a conditional generator. This is a training-by-sampling method used to access all classes. During the process, data from each class are evenly resampled from discrete columns but may not have a uniform distribution. This guarantees a class balance during the training process [45].

After a dataset is synthesized, it is necessary to evaluate its statistical characteristics to ensure usability. This can be performed using correlation analysis among the variables of the real and synthetic datasets. If they are identical, the correlation matrix should be the same. The Pearson correlation coefficient, r, is a common matrix that indicates the relationship between two variables and is defined as

$$r = \frac{\Sigma_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{N}(x_i - \bar{x})^2 \Sigma_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{9}$$

where x_i and y_i are random samples indexed with i, N is sample size, and $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively. Other matrices for evaluation include the central tendency, PCA, and some privacy perspectives, such as identical data points between real and fake datasets or the mean and standard deviation of the distance between each fake record and the most identical real records [46].

The Kolmogorov–Smirnov (KS) test is a statistical test used to evaluate similarity in terms of the distribution between a sample and a reference cumulative probability distribution of an attribute [47], [48]. Given N ordered data points x_1, x_2, ..., x_N, the empirical cumulative distribution function (CDF), $F_s(x)$, is defined as a step function that increases by $1/N$ at each observation point x_n:

$$F_s(x) = \begin{cases} 0 & x < x_1 \\ \frac{k}{N} & x_n \leq x < x_{n+1} \\ 1 & x \geq x_N, \end{cases} \tag{10}$$

where k is the number of instances less than or equal to x and $n \in \{1, \ldots, N\}$. Note that the CDF of a discrete variable is a step function, and $F_s(x)$ is an estimator of the true CDF based on a sample of observations, which we can show that $F_s(x) = F_s(x_n)$.

The KS statistic, D, for a reference cumulative probability distribution, $F_r(x_n)$, is defined as

$$D = \max_{1 \leq n \leq N} |F_s(x_n) - F_r(x_n)|. \tag{11}$$

When the CDF of an observed sample, $F_s(x_n)$, is close to $F_r(x_n)$, then D approaches zero [49].

Furthermore, if we compare similar attributes from two datasets, $F_u(x_n)$ and $F_v(x_n)$, we can determine whether they represent the same distribution. Thus, the KS statistic used under these circumstances is the "Two-Sample KS test." The KS statistic, $D_{u,v}$, is defined as

$$D_{u,v} = \max_{1 \leq n \leq N} |F_u(x_n) - F_v(x_n)|, \tag{12}$$

where $F_u(x_n)$ and $F_v(x_n)$ are the CDF of the datasets u and v, respectively. Similarly, samples $F_u(x_n)$ and $F_v(x_n)$ represent similar distributions when $D_{u,v}$ converges to zero.

To obtain the KS statistic, we applied the "Quality Report" function from the SDMetrics [50]. Here, KS statistics of every attribute in a dataset are computed. Subsequently, the Quality Score is computed, and represents the average KS statistics of all attributes. Note that, in this package, the Quality Score is reported as $1 - D_{u,v}$. Thus, for evaluation, the higher the score, the higher the quality.

2) RESAMPLING TECHNIQUE

Resampling techniques are popular approaches to address an imbalanced dataset. Undersampling eliminates a major class to decrease the number of instances in that class, whereas oversampling increases the minor class to balance the overall dataset [51], [52], [53]. These techniques do not require any synthetic processes. However, with oversampling, some drawbacks such as long training times and duplication of instances should be considered. Thus, SMOTE is a possible oversampling technique that can synthesize extra minority class instances based on k-nearest neighbors and mitigate these drawbacks [54].

In this subsection, an overview of the method for augmenting or synthesizing a dataset from a real dataset with limited instances for sufficient instances and evaluation methods is presented. The supervised learning is described in the following subsections.

E. SUPERVISED LEARNING

In this subsection, we briefly describe supervised learning, particularly the classification models, and their evaluation metrics. These models are used in our work.

In our work, we used classification models, such as Decision Tree, k-nearest neighbor (K-NN), Logistic Regression, and Neural Networks, to prove the usability of the synthetic datasets. These classification models are categorized as supervised learning techniques, where categorized (or nominal) labels are necessary. Another model is the regression model, in which the labels are numerical (discrete or continuous). This model is not utilized in our work because we considered the categorized labels. To evaluate the performance of classification models, the matrices below are considered [55].

- Accuracy indicates the rate at which a model provides correct predictions.
- Precision indicates the successful detection of "true positives" in the positive labels.
- Recall indicates the successful detection of "true positives" in the positive samples.
- If we consider the relationship between Precision and Recall, where the weights of both metrics are equal, the F1-Score can be used to evaluate this relationship.

In addition, k-fold cross-validation can use to evaluate. It divides a dataset into k subsets (or folds) of equal size. The k - 1 subset is used to train a model, and the remaining subset is used to test the model [55]. The cross-validation supports the evaluation phase when a number of instances is limited. 10-folds cross-validation is a standard number of folds. Related works are presented in the following subsections.

F. RELATED WORK

First, GAN is applied to support the oversampling technique and increase the minority class of an imbalanced dataset [56]. In the experiment, the credit card dataset had two classes: fraudulent and nonfraudulent. Fraudulent instances are oversampled through GAN to obtain more instances. In addition, GAN is used to detect fraud in credit card usage [57], develop DUO GAN to synthesize heavily unbalanced data [58], or ensemble GAN to synthesize data [59].

Moreover, TGAN is developed to support tabular data synthesis [44]. It is tested by comparing the efficacy of the machine learning models between real and synthetic tabular datasets. TGAN is used to generate additional data, and is measured using the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to compare the results [46].

Furthermore, CTGAN is developed to support the generation of tabular data using training-by-sampling to ensure balance between classes [45]. CTGAN is adopted in other studies for the synthesis of datasets. It is used to generate additional custom duty instances for the analysis [60]. This algorithm is also used to synthesize additional data to train an intrusion detection system for cybersecurity tasks [61]. Moreover, it supported the estimation of the required electrical power by generating additional data for analysis [62].

Machine learning has also been used to detect tax frauds. Supervised learning has been used to detect taxpayers using false invoicing to evade taxes [63], identify suspicious tax evasion groups [33], and boost their performance in detecting tax evasion [9]. In addition, neural networks has been applied to enhance the efficiency of machine learning, such as for the detection of fraud in tax declarations using Ensemble ISGNN [64], Personal Income Tax evasion [65], and transaction-based tax evasion [66]. Unsupervised learning has been applied to detect tax fraud using association rules and dimensionality reduction [67], to identify underreporting declarations using the clustering approach [68], and to detect VAT fraud by computing the anomaly score using Fixed-Width Anomaly Detection (FWAD) and Local Outlier Factor (LOF) [69]. Furthermore, several complex algorithms have been developed to efficiently detect tax evasion cases, such as Hybrid Unsupervised Outlier Detection (HUNOD) using K-means clustering, autoencoder-based outlier detection [12], and unsupervised conditional adversarial networks (UCAN) using a generative model to detect tax evasion [70]. Finally, a synthetic dataset is used to analyze bank transactions for the normal and suspicious behaviors of shell companies using a Banking Transaction Simulator or BTS [29]. In this experiment, two sets of data are synthesized: transactions of normal and suspicious cases. Subsequently, instances from the synthetic dataset are assigned a LOF score.

III. PROPOSED PROCESS

The processes in this work are divided into three main parts, as shown in Fig. 1. First, Data Acquisition is a step to obtain a dataset. The financial statements listed in SET [4] are selected as the main source. Second, Data Augmentation is a step to increase the number of instances using CTGAN and SMOTE. Finally, Model Development is a step to build a model using the synthetic datasets as inputs.

A. DATA ACQUISITION

Data acquisition is the first step in obtaining the dataset. Four steps are involved in this process: Data Gathering, Data Filtering, Attribute Generation, and Stratified Sampling.

First of all, the first step is "Data Gathering." We begin by exploring tax return forms as well as risk criteria. This provides information on the required attributes. Based on our scope, the consideration of CIT and VAT is our task in this work. Hence, the CIT 50 and PP30 are studied. Such forms are available on the website of the Revenue Department [71]. Simultaneously, the risk assessment criteria are studied to select criteria for our work. Not all the criteria presented in Table 1 are considered. We choose criteria whose required attributes can be retrieved directly from the financial statements. In addition, the selected criteria should relate to VAT and CIT, which is indicated in our scope of work, especially in retail or wholesale businesses, as the attributes are obvious. For the first reason, criteria 1, 6, and 7 in Table 1 are selected. Second, although E_p cannot be directly retrieved from financial statements, it relates to retail and wholesale businesses. Consequently, criterion 8 from Table 1 is selected for analysis. These criteria have different weighted scores that affect outputs and create more complications. Therefore, equivalent weighted scores of the selected criteria are assumed to simplify this work.

With the knowledge of the selected criteria and required attributes, we gather data by accessing an annual report from an open source. Annual reports with required features are retrieved from the SET website [4]. For our work, data for five years from 2017 to 2021 are collected and combined into a single dataset. To simplify this process, a spreadsheet is recommended.

Secondly, after we collect and obtain a dataset, we need to perform "Data Filtering" to filter out inappropriate values. Under these circumstances, samples that either contain missing values or do not represent the actual behavior of the dataset, such as negative or zero values, are removed.

Third, as some required attributes for analysis are not available in the financial statement, we perform "Attribute Generation." At this point, we apply financial equations to obtain them. In addition, labels are assigned based on the selected criteria. If any samples are subjected to the selected risk assessment criteria for at least two criteria, they are labeled as "1," otherwise labeled "0." Label 1 is assigned based on the selected criteria, indicating the need for further analysis. It is declared to be a suspicious case rather than an evasive one. For example, the reason for being labeled 1 may come from economic situations that make a company subject to the Gross Loss criterion. Therefore, a further audit is compulsory before deciding whether an instance with the

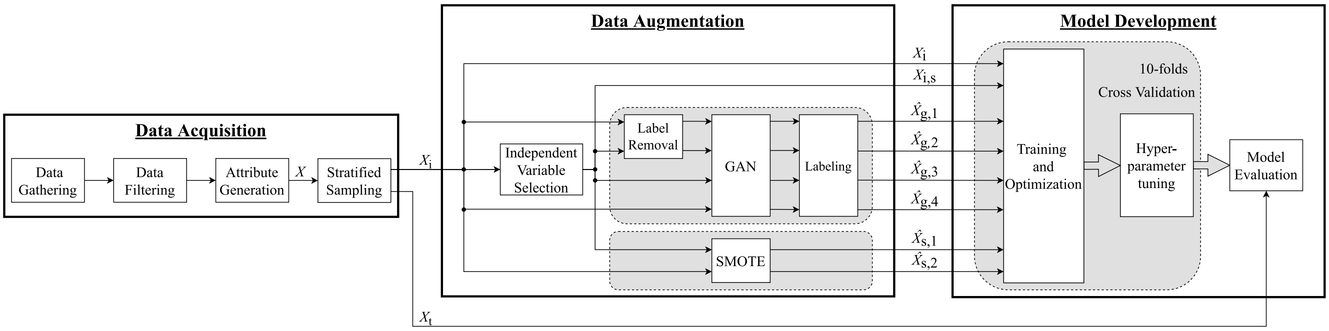


FIGURE 1. The proposed process.

TABLE 3. An example of a dataset with independent variables.

R_s	I_b	I_e	E_c	**Label**
75538	3718743	889684	211316	**1**
3045093	37223	44647	41466	**0**
27179177	43284336	22068084	23613834	**0**

TABLE 4. An example of a dataset with all variables.

R_s	I_b	E_p	I_s	I_e	E_c	P_g	**Label**
20619	878	17811	18690	749	17941	2677	**0**
2064	514	2876	3390	5949	2795	-731	**1**
21275	164	15589	157539	1828	15571	5704	**0**

label 1 evades tax. Labels are assigned in this step, because this is a requirement for stratified sampling and SMOTE. After completing these steps, the collected dataset is obtained and is defined as X in Fig. 1. For simplicity, the formulae and selected risk criteria can be applied to a spreadsheet to generate required attributes.

Furthermore, if financial statements are downloaded from other stock market repositories, the tax return datasets can be obtained by applying these three steps. Tax return forms should be retrieved from a particular region to obtain a dataset that demonstrates the behavior of financial data in that region.

Finally, X undergoes "Stratified Sampling" process. Stratified sampling is applied to guarantee distribution between classes. In this step, X is split into two equal sets using stratified sampling: X_i and X_t. X_i is used to train models as well as to augment to increase the volume of the data. X_t is used as the testing dataset for the model development process.

The steps for constructing and preparing the datasets have now been completed. In the next subsection, the augmentation of the obtained dataset for more instances is discussed.

B. DATA AUGMENTATION

In the previous subsection, the steps to construct a dataset are discussed to demonstrate the method used to obtain the datasets. Now, we are proceeding to the second process "Data Augmentation," where X_i in Fig. 1 is augmented.

First, the attributes in X_i are categorized as independent and dependent variables. The independent variables are those that are not dependent on other variables, and do not change as the setting of the experiment is changed. Dependent variables are those that are dependent on other variables and vary as independent or dependent variables are changed. Under these circumstances, the independent variables are directly collected from the financial statements, as shown in Table 3. Unlike independent variables, dependent variables may not need to be retrieved directly from financial statements, as shown in Table 4. They may either appear in financial statements such as P_g or require further calculations based on financial equations such as E_p.

Subsequently, as the independent and dependent variables for our work are defined, the "Independent Variable Selection" step shown in Fig. 1 is the next step. In this step, four datasets are provided with only independent variables, whereas the remaining datasets are provided with both the dependent and independent variables. One dataset is defined as $X_\mathrm{i,s}$ in Fig. 1, which is a collected dataset with only independent variables. This dataset is used for training the models. The remaining datasets are augmented in this step. Two datasets are inputs for GAN, and the another is for SMOTE. The purpose of this step is to compare the results of the input variations.

In addition, two of the four datasets augmented using GAN undergoes a "Label Removal" step, as shown in Fig. 1. This can differentiate between datasets synthesized with and without labels.

Consequently, the inputs for GAN are as follows:

- A dataset with independent and dependent variables and no labels.
- A dataset with only independent variables and no labels.
- A dataset with only independent variables and labels.
- A dataset with independent and dependent variables and labels.

Subsequently, X_i is augmented for more instances using CTGAN. CTGAN is chosen for our work because one advantage is training-by-sampling, which considers class distribution [45]. It is available at the sdv-dev GitHub and can be installed using the "synthetic data vault" package [72], [73].

For our work, we run the algorithm on the Jupyter Notebook using the Google Collab platform.

Additionally, the SMOTE technique is applied to increase the number of instances. However, SMOTE requires labels. Therefore, the "Label Removal" step in Fig. 1 does not apply. Although the concept of synthesis is different from that of CTGAN, which captures the relationship between the original data and synthesizes a new sample using a neural network [45], SMOTE oversamples the minor class and synthesis based on the k-nearest neighbor [54]. In our experiment, we apply SMOTE to compare the results with those obtained using CTGAN.

After we obtain the synthetic datasets from this process, all the synthetic datasets from CTGAN are assigned labels based on the selected risk criteria. We do not plan to measure whether CTGAN provides accurate results. We only want to observe the influence of labels on the other attributes. This is why labels from GAN are removed, and new labels are allocated regardless of their appearance prior to GAN. This step is shown as "Labeling" in Fig. 1. However, SMOTE does not require this step because all labels generate during the stratified sampling step are already correct.

Thus, we obtain six synthetic datasets, as shown in Fig. 1, which are the outputs of the Data Augmentation process. The synthetic datasets are as follows:

- $\hat{X}_{\mathrm{g},1}$ is a synthetic dataset generated using CTGAN from an input with independent and dependent variables and no labels.
- $\hat{X}_{\mathrm{g},2}$ is a synthetic dataset generated using CTGAN from an input with only independent variables and no labels.
- $\hat{X}_{\mathrm{g},3}$ is a synthetic dataset generated using CTGAN from an input with only independent variables and labels.
- $\hat{X}_{\mathrm{g},4}$ is a synthetic dataset generated using CTGAN from an input with independent and dependent variables and labels.
- $\hat{X}_{\mathrm{s},1}$ is a synthetic dataset generated using SMOTE from an input with only independent variables and labels.
- $\hat{X}_{\mathrm{s},2}$ is a synthetic dataset generated using SMOTE from an input with independent and dependent variables and labels.

During this process, X_i is synthesized and yield six datasets. This indicates that we complete the "Data Acquisition" and "Data Augmentation" processes.

C. MODEL DEVELOPMENT

After the "Data Acquisition" and "Data Augmentation" steps are completed, the collected and synthetic datasets are used as inputs in the third process "Model Development." The training sets for this step are X_i, $X_{\mathrm{i,s}}$, $\hat{X}_{\mathrm{g},1}$, $\hat{X}_{\mathrm{g},2}$, $\hat{X}_{\mathrm{g},3}$, $\hat{X}_{\mathrm{g},4}$, $\hat{X}_{\mathrm{s},1}$, and $\hat{X}_{\mathrm{s},2}$. The testing set is X_t. In this step, the software "RapidMiner Studio" software [74] is used to tune hyperparameters and construct machine learning models.

After the training datasets are obtained, they are used as inputs for cross-validation. As shown in Fig. 1, we use 10-fold cross-validation. During "Training and Optimization," hyperparameter-tuning is performed to figure out hyperparameters of the selected models. These are Decision Trees, K-NN, Logistic Regression, and Neural Networks. The first three of the selected models are fundamental models, used to illustrate the baseline performance. The baseline performance is considered as the minimum expected performance of a model [75]. In addition, fundamental models typically have fewer hyperparameters than complex models. We also apply Neural Networks to compare the results with the fundamental ones. During the testing phase of cross-validation, the "hyperparameter tuning" is performed. Consequently, the hyperparameters for model evaluation are the outputs of the cross-validation step.

As the hyperparameters for the selected models are obtained, we continue with "Model Evaluation." The training sets are the collected and synthetic datasets obtained from the previous steps and the hyperparameters of the selected models obtain from hyperparameter tuning are utilized. The testing set is X_t. Consequently, the performance is calculated based on the matrices mentioned in Section II-E, with scores ranging from zero to one, where zero is the worst score and one is the perfect score. From this index, we obtain the output of this process, which is a selected model.

This section discusses the proposed process. These processes include generating datasets, augmenting them for additional instances, and selecting an appropriate model. In the next section, results are presented and discussed.

IV. EVALUATION RESULT AND DISCUSSION

In the previous section, we describe the process beginning with obtaining a dataset, followed by augmenting it, and finally selecting an appropriate model. This section presents and discusses the results obtained by completing the proposed process.

A. A COLLECTED DATASET

The number of instances collected from 2017 to 2021, after the data acquisition process is completed, is listed in Table 5. From Table 5, the total number of collected instances is 3,852, where 3,105 instances are completed data and 747 instances are missing data. From the completed data, 2,675 are labeled as "0" and 430 are labeled as "1." This indicates that about one-sixth of the collected data are labeled. Thus, the collected data is imbalanced. Subsequently, data filtering is performed to remove unreasonable values. Under our circumstances, an unreasonable value is defined as a value that does not reflect reality, such as inventory attributes less than zero or incomplete values. Thus, the total number of appropriate instances for further analysis is 2,942 instances. X is split in half for synthesis. Thus, the total number of instances for "Data Augmentation" is 1,471. Generally, it is sufficient to build fundamental models with 1,000 instances; however, for complex models, such as artificial neural networks (ANNs) or deep learning, more instances are required [76]. Thus, we conduct this work to demonstrate that data can be synthesized to access more instances for an analysis.

Several remarks are made after collecting the dataset. First, the units in financial statements should be carefully read. Some businesses report values in Baht, Thousand Baht, and Million Baht. It may also appear in foreign currencies such as the US dollar. Second, when downloading financial statements, several attributes may not appear in the Profit and Loss section, but, instead, in the asset section. For example, I_b or I_e may appear in the asset section. P_g may not appear in financial statements; thus, the "Attribute Generation" in Fig. 1 is applied by computing (4). Next, it should be emphasized that E_c refers only to what is paid for goods and services, such as the cost of goods or services. This excludes all operating expenses and financial costs. Finally, other attributes that are not mentioned can be obtained by applying the financial equations as long as they are not independent variables, as presented in Table 3.

Fig. 2 presents a scatter matrix of X. Only the independent variables R_s, I_b, I_e, and E_c are presented. These attributes are collected directly from the financial statements. The four attributes of X are paired to create a scatter plot. The scatter plots can indicate the relationship between attributes. For example, from the scatter plot of R_s and E_c in the lower left corner of Fig. 2, it can be inferred that R_s and E_c have a strong positive correlation. The level of correlation can be used for attribute selection during model development. In each plot, the blue and orange colors indicate the labels "0" and "1," respectively. These labels are given based on the selected criteria. Considering the labels, there are some pairs of attributes that can separate the areas of labels 0 and 1, such as E_c and I_b, or E_c and I_e. Based on these relations, we can build a machine learning model to classify labels. The histograms of each attribute are presented as diagonal plots in Fig. 2. Each histogram has two lines, where the blue and orange lines represent the labels 0 and 1, respectively. We can see that the histogram of label 0 is bigger than that of label 1. This is because the instance number with label 0 is much greater than that with label 1.

To complete a stratified sampling step, labels are required. As we conduct this research to detect tax evasion, we only access public financial data from SET and rule-based criteria presented in Table 1. Thus, we begin our research with these data and criteria. They may be complex as we marked an instance with "1" when it corresponds to at least two criteria. As mentioned earlier, the label given in our study indicates only suspicious behavior not evaded behavior. In addition, the data from these sources are audited by specialists. Thus, they are not actually an evaded case. But, tax data are sensitive and not disclosed to the public, by starting with our proposed idea, we can obtain data for training a machine. Then, a machine will learn from these inputs and adapt to new unseen behavior, as tax fraudsters change their suspicious behavior over time [14]. Gradually, we will be able to explore new behavior for tax-related research. In the future, when a machine becomes more advanced, a multi-class label can be applied to provide new insights. For example, a label that represents different risk levels or provides a potential risk description of each instance can be developed. This development will benefit officials in selecting a potential company for auditing and solving definite tax-fraud-related issues.

TABLE 5. The collected dataset.

Status/Year	2021	2020	2019	2018	2017	Total
Completed	689	647	619	587	563	3,105
Missing	149	154	146	148	150	747
Yearly Total	838	801	765	735	713	3,852

B. AN AUGMENTED DATASET

In our work, we apply 40,000 epochs and 250 batch sizes to train the CTGAN to generate the synthetic datasets. The total number of synthetic instances from CTGAN and SMOTE for each dataset is 5,000. After obtaining the synthetic datasets, the correlation matrix, PCA scatter plot, and Quality Score [50] of these datasets are used to evaluate synthesis quality. The correlation matrices for $X_{i,s}$, $\hat{X}_{g,2}$, $\hat{X}_{g,3}$, and $\hat{X}_{s,1}$ are presented in Figs. 3a, 3b, 3c, and 3d, respectively. The red and blue colors represent positive and negative values of r, respectively. The shade of the color depends on the magnitude of r. For example, in Fig. 3a, the correlation between E_c and R_s is shown in darker red. This indicates a strong correlation between the two variables. However, considering the correlation matrix in Figs. 3a, 3c, and 3d, r of the label rows and columns is close to zero, which represents a weak or no correlation. If the synthetic and original datasets represent the same relationship, then the correlation matrix should be as similar as possible. Moreover, by comparing Figs. 3a and 3b, the relationship between the attributes in $\hat{X}_{g,2}$ matches that in $X_{i,s}$. An example is the relationship between E_c and R_s of $X_{i,s}$ in Fig. 3a and that of $\hat{X}_{g,2}$ in Fig. 3b. We can see that r of E_c and R_s are 0.99 and 0.83 for $X_{i,s}$ and $\hat{X}_{g,2}$, respectively. Fig. 3b shows the correlation matrix of a synthetic dataset that does not obtain label attributes prior to CTGAN. Thus, label attributes prior to CTGAN do not influence the augmentation process. In general, most correlations are visible for other attributes and indicate a positive relationship as r is greater than 0 and close to 1.

Another important property of data is the distribution in which a scatter plot can be demonstrated. However, it is limited to low-dimensional data such as two or three attributes. Therefore, PCA is applied to overcome this limitation. The PCA scatter plots of $X_{i,s}$, $\hat{X}_{g,2}$, $\hat{X}_{g,3}$, and $\hat{X}_{s,1}$ are shown in Figs. 4a, 4b, 4c, and 4d, respectively. Each dataset undergoes PCA to reduce its dimensions, which are the attributes of the data, to two dimensions: pc_1 and pc_2. Fig. 4b shows a PCA of $\hat{X}_{g,2}$, which is different from the others because the other PCA scatter plots have the label attribute included in the PCA process, while $\hat{X}_{g,2}$ does not. If two datasets have the same statistical characteristics, the PCA scatter plots should be similar. From this evaluation, it can be observed that Figs. 4a and 4c are similar. This implies that the distributions of these datasets are the same.

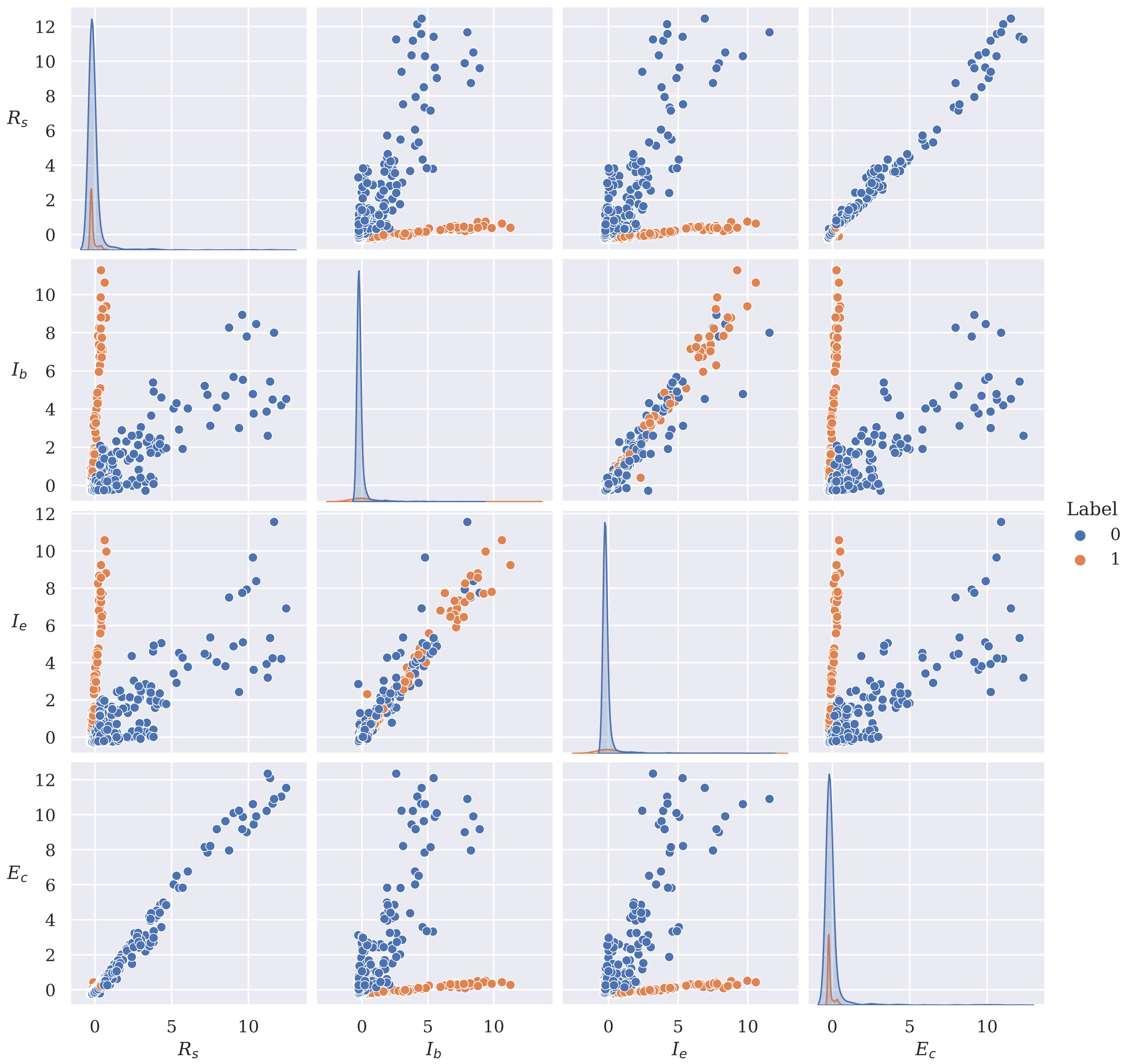


FIGURE 2. An example of Scatter Plot Matrix of X where only independent variables are included.

TABLE 6. The Quality Score the synthetic datasets.

Technique	Synthetic Dataset	Quality Score
CTGAN	$\hat{X}_{g,1}$	0.859
	$\hat{X}_{g,2}$	0.853
	$\hat{X}_{g,3}$	0.850
	$\hat{X}_{g,4}$	0.856
SMOTE	$\hat{X}_{s,1}$	0.894
	$\hat{X}_{s,2}$	0.895

Next, the Quality Scores of the synthetic datasets are presented. As mentioned in Section II-D1, the higher the score, the higher the quality. Table 6 lists the Quality Scores of the synthetic datasets, with an average Quality Score of 0.868. CTGAN yields an average Quality Score of 0.855, whereas SMOTE yields the highest score of 0.895. With CTGAN, $\hat{X}_{g,1}$ provides the highest Quality Score at 0.859, whereas the others have also high scores that do not deviate significantly from $\hat{X}_{g,1}$. With SMOTE, both datasets yields similar scores of approximately 0.895. The results show that SMOTE, has a better overall score than that of CTGAN. The reason is that SMOTE uses nearest-neighbor techniques to increase the number of minorities [54]. In contrast, CTGAN generates instances based on relationships and does not access the original dataset [43]. Thus, the synthetic instances from SMOTE are more similar to the original instances than those from CTGAN.

Considering the synthetic datasets without labels ($\hat{X}_{g,1}$ and $\hat{X}_{g,2}$) and with labels ($\hat{X}_{g,3}$ and $\hat{X}_{g,4}$), the average Quality Scores are 0.856 and 0.853, respectively. Thus, the label does not have any influence on the synthesis. Considering $\hat{X}_{g,1}$ and $\hat{X}_{g,4}$, which contain the dependent and independent variables, the average Quality Score is 0.858. For $\hat{X}_{g,2}$ and $\hat{X}_{g,3}$, which contain only independent variables, the Quality Score is 0.852. Synthesizing either with only independent variables or both dependent and independent variables do not indicate any notable differences. However, no obvious differentiation can be observed after the synthesis using SMOTE. In conclusion, the label does not influence the data augmentation process, and only the independent variables provide sufficient information for synthesis.

The data synthesizing techniques available include adding noise to a dataset to generate new unique values, applying the Bayesian method to generate new instances, or creating new data based on criteria obtained from Classification and Regression Tree (CART) [77]. These methods are simpler compared to the GAN technique utilized in our research. Our collected dataset is a simple structured dataset, as we only select relevant variables. For instance, considering a CIT tax return form [71], enterprises are required to fill in only the necessary information that fits their specific circumstances. This results in different sets of required data for each enterprise. Constructing a synthetic dataset from this type of data can be challenging since the data is unstructured. GAN is one of the suitable synthesizing algorithms that considers the relationship among attributes and is appropriate for generating synthetic datasets. To demonstrate the potential of GAN in synthesizing datasets for further analysis, we have taken a first step by using a simple dataset. By increasing the number of instances, this technique could be used to construct more datasets in this area.

In this step, we obtain the synthetic datasets generated either with or without labels, and contained either only independent variables or dependent and independent variables. Several evaluation methods, including the correlation matrix, PCA scatter plot, and Quality Score, are used to assess the similarity between the original and synthetic datasets. In the next subsection, machine learning models of tax evasion detection are presented.

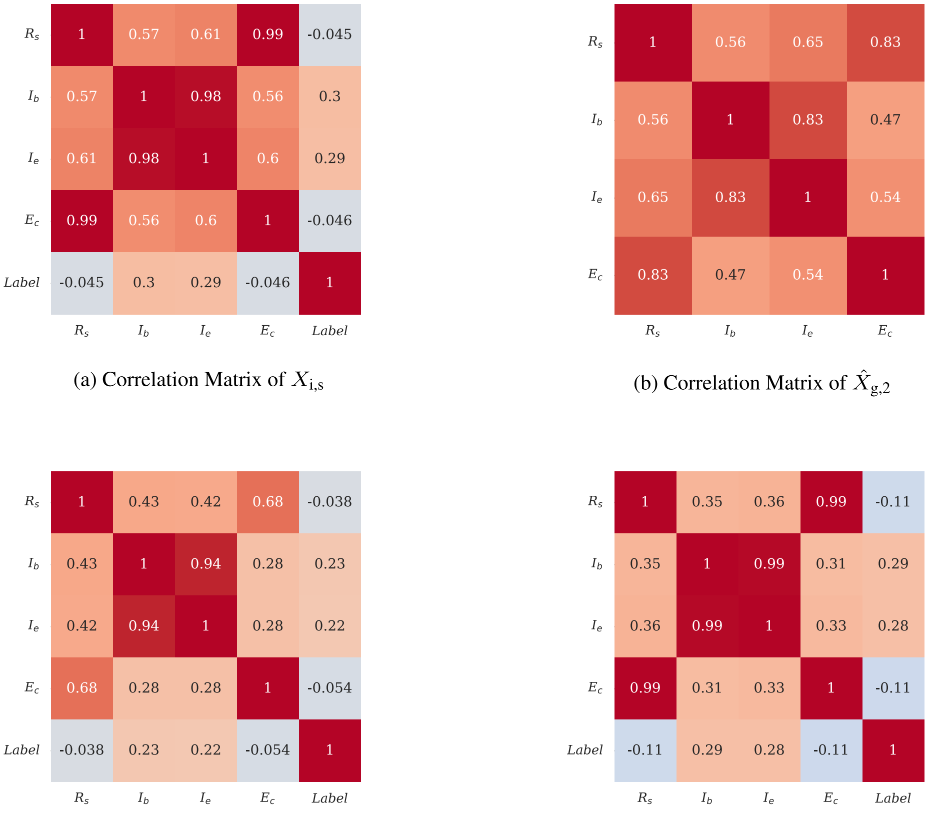
C. PERFORMANCES OF TRAINED MODELS

Four supervised learning models are developed by training the models with the collected and the synthetic datasets. In addition, hyperparameter tuning is performed using grid optimization on the same datasets. At this step, as mentioned in Section III-C, we apply the machine learning models to evaluate baseline performance. Subsequently, the performance of the models is evaluated using the tested set, X_t, which is an actual dataset. Precision and Recall evaluated in this Section represent the performance in case the models detect instances with the label "1" correctly. The results of the evaluation are as follows.

Table 7 shows the performance of the models trained with the collected dataset, and tested using X_t. There are two types of inputs: X_i and $X_{i,s}$. The former contains all collected variables shown in Table 4, and the latter contains only independent variables shown in Table 3. According to Table 7, the average Accuracy, Precision, Recall, and F1-Score of X_i are 0.96, 0.89, 0.64, and 0.73, respectively. The average Accuracy, Precision, Recall, and F1-Score of $X_{i,s}$ are 0.95, 0.83, 0.66, and 0.73, respectively. Generally, models trained with only independent variables yield similar performances to those trained with all variables. First, considering the Decision Tree, when it is trained with X_i, it yields better performance than trained with $X_{i,s}$. As the collected dataset is imbalanced, with only one-sixth of all instances labeled, this could potentially impact the performance of our machine learning models. However, we have decided to leave this issue unaddressed. This is because if we attempt to fix the problem through down-sampling or up-sampling, it could significantly alter the behavior of our training data. Second, for K-NN, the performance is the highest among the others. Under our circumstances, the number of nearest neighbors, k, is five, and the Euclidean Distance is used to evaluate the similarity. Third, Logistic Regression yields high performance. As we test our dataset with this model, the stage of our model is high variance or overfitting [55]. This means more data can alleviate this issue. Lastly, Neural Network is applied. The given Neural Network has a structure consisting of five layers, with each layer containing six nodes, and the activation function is the sigmoid function. In this case, the Neural Network is trained for 1980 training cycles, which can be interpreted as 1980 epochs. At the end of each epoch, the weights of each connection in the network are adjusted to minimize the error function. In terms of performance, the Accuracy of Neural Networks trained with both datasets is similar. However, Neural Network, using $X_{i,s}$ as a training set, provides a slightly better F1-Score than using X_i as a training set. Considering Recall, the results are the lowest among the results from other models. Moreover, Neural Networks can also have high variances. Thus, more samples can mitigate this issue.

Table 8 shows the performance of the models trained using the synthetic datasets from CTGAN. The average Accuracy, Precision, Recall, and F1-Score are 0.94, 0.90, 0.84, and 0.87, respectively. K-NN yields the best performance with an average Accuracy of 0.96 and an average F1-Score of 0.93. Decision Tree yields an average Accuracy of 0.93, and an average F1-Score of 0.87.

Based on our results, we observe that the models trained on $\hat{X}_{g,3}$ generally perform better than those trained on $\hat{X}_{g,4}$. This is due to the fact that the models trained on $\hat{X}_{g,3}$ use fewer features, which can result in better performance compared to models trained using all features. Selecting important features during model development is another factor that can enhance the model's performance [78], [79]. However, our research do not focus on feature selection to obtain the best model, so it is possible that the model using only necessary features



(a) Correlation Matrix of $X_{\text{i,s}}$

(b) Correlation Matrix of $\hat{X}_{\text{g,2}}$

(c) Correlation Matrix of $\hat{X}_{\text{g,3}}$

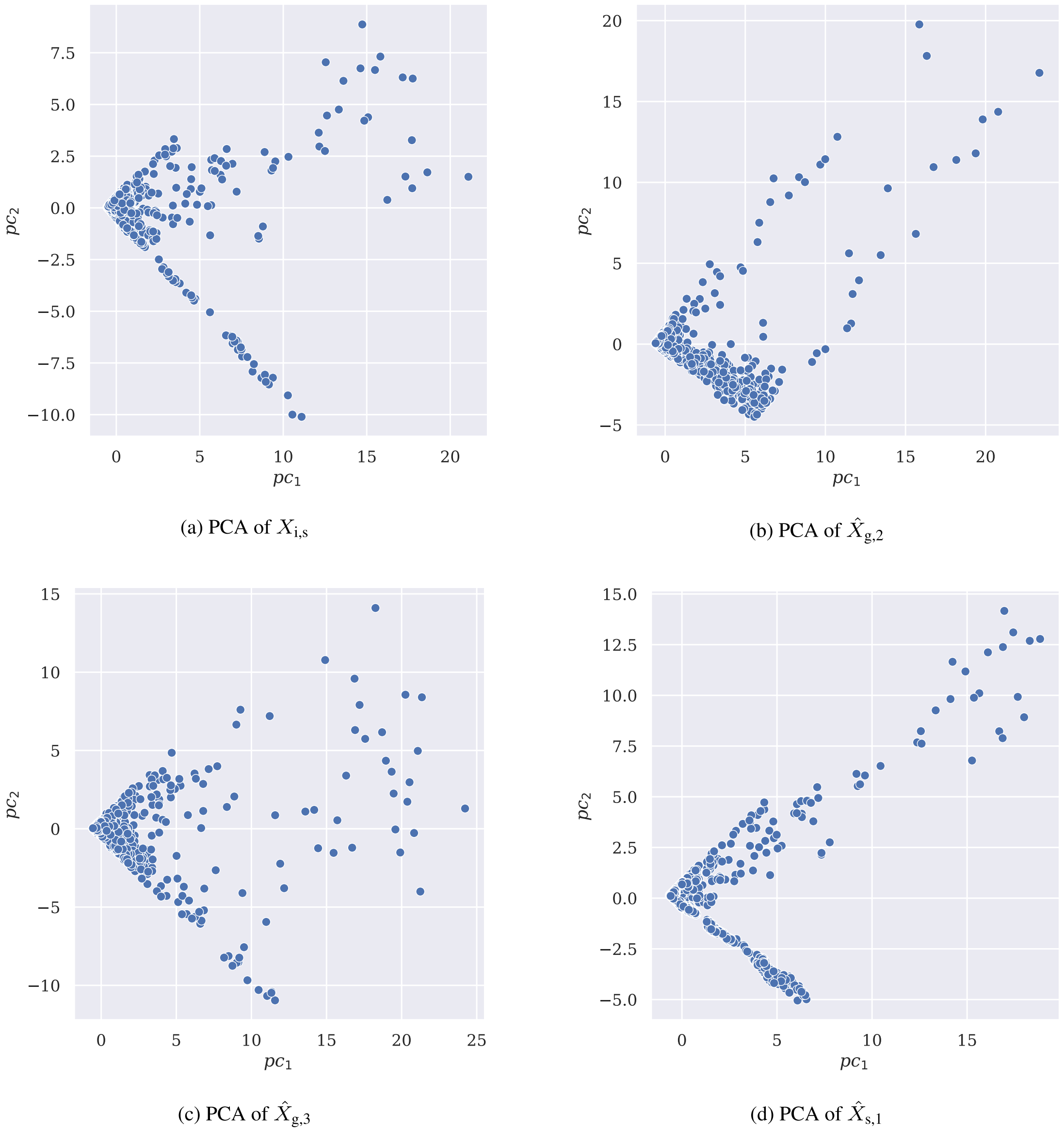
(d) Correlation Matrix of $\hat{X}_{\text{s,1}}$

FIGURE 3. Correlation Matrices of the selected datasets.

performs better than the model using all features, which may lead to overfitting. However, we evaluate the baseline performance of each model to see the minimum possible performance of each model trained with different synthetic datasets. Logistic Regression yields an average Accuracy of 0.91, and an average F1-Score of 0.87. Neural Networks yields satisfactory performance. Its average Accuracy and average F1-Score are 0.94, and 0.80, respectively. In short, K-NN provides the highest performance, its performance does not notably different from the other models. These algorithms perform well with the training datasets that are generated by CTGAN.

Compared with the models trained with the collected dataset, the synthetic datasets using CTGAN have slightly better performance. Decision Tree trained with the collected dataset performs better than the synthetic ones when considering an Accuracy. But, as we generate more instances with CTGAN, the F1-Score is increase. This indicates that the Precision and Recall of models trained with the synthetic datasets are higher than the ones trained with the collected dataset. Similarly, K-NN and Logistics Regression trained with the synthetic datasets using CTGAN yield lower Accuracy, while the F1-Score increases. Moreover, Neural Networks trained with the synthetic datasets using CTGAN yields similar Accuracy to those trained with the collected dataset. However, the F1-Score of Neural Networks trained with the synthetic datasets is improved. With complex models, more parameters, such as hidden layers and their nodes, need to be analyzed compared to the fundamental ones. Generally, one plausible factor for increasing in F1-Score when training with the synthetic datasets using CTGAN is the number of training samples in the training set. The collected dataset contains 1,471 instances whereas the synthetics datasets contain 5,000 samples. This factor can improve the performance of these models.

Table 9 shows the results of the models trained using the synthetic data from SMOTE, where k is the number of nearest neighbors. The average Accuracy, Precision, Recall, and F1-Score of the training datasets oversampled using SMOTE are 0.96, 0.97, 0.95, and 0.96, respectively. The algorithm that yields the best performance is K-NN, which provides an average Accuracy of 0.98 and an average F1-Score of 0.98. Decision Tree and Logistic Regression yield average Accuracies of 0.97 and 0.95, respectively, and average F1-Score of 0.97 and 0.95, respectively. Neural Network yields a good performance. Its average Accuracy and average F1-Score are 0.92, and 0.92, respectively. When we change the number of nearest neighbors to either three or five, the performances of the models remain relatively the same. With three nearest neighbor, both average Accuracy of $\hat{X}_{\text{s,1}}$ and



(a) PCA of $X_{\mathrm{i,s}}$

(b) PCA of $\hat{X}_{\mathrm{g},2}$

(c) PCA of $\hat{X}_{\mathrm{g},3}$

(d) PCA of $\hat{X}_{\mathrm{s},1}$

FIGURE 4. **PCA scatter plots of the selected datasets.**

$\hat{X}_{\mathrm{s},2}$ is approximately 0.96. With five nearest neighbor, both average Accuracy of $\hat{X}_{\mathrm{s},1}$ and $\hat{X}_{\mathrm{s},2}$ is approximately 0.96. Therefore, significant differences are not visible.

Compared with the models trained with the collected dataset, the synthetic datasets using SMOTE have higher performance. Decision Tree trained with synthetic datasets from SMOTE yields higher performance than the one trained with X_{i} and $X_{\mathrm{i,s}}$. This occurs because the data points generated from SMOTE are stuck densely together as SMOTE synthesizes based on its nearest neighbors. Thus, the synthetic datasets using SMOTE may reflect a good relationship among features and, thus, yield higher results. With K-NN and Logistic Regression, the performances are also improved. Emphatically, one big improvement is the performance of Neural Networks between the models that trained with the collected datasets and the synthetic datasets using SMOTE. With the Neural Network, the Accuracy of a model trained with either $\hat{X}_{\mathrm{s},1}$ or $\hat{X}_{\mathrm{s},2}$ is similar to those of X_{i} and $X_{\mathrm{i,s}}$ while F1-Score are higher than those of X_{i} and $X_{\mathrm{i,s}}$. Compared with Table 8, the SMOTEs' average Accuracy of all algorithms, which is approximately 0.96, is slightly better than that of CTGANs'. With synthetic datasets that are generated from either CTGAN or SMOTE and have a similar relationship among features and similar distribution to the collected dataset, the model can learn and provide predictions.

Finally, from a theoretical perspective, the processes of synthesizing data using CTGAN and SMOTE are different. From the results, it cannot be concluded which technique, CTGAN or SMOTE, provides the best results. It yields results that cannot distinguish from one another. CTGAN uses a conditional GAN to learn from the original data or training-by-sampling [45], whereas SMOTE uses k-nearest

TABLE 7. Performance of models using the collected datasets as training sets.

Dataset	Algorithm	Accuracy	Precision	Recall	F1-Score
X_{i}	**Decision Tree**	0.962	0.852	0.677	0.754
	K-NN	0.978	0.928	0.811	0.866
	Logistic Regression	0.959	0.853	0.638	0.730
	Neural Networks	0.948	0.932	0.431	0.589
$X_{\text{i,s}}$	**Decision Tree**	0.948	0.734	0.686	0.709
	K-NN	0.971	0.913	0.766	0.833
	Logistic Regression	0.958	0.871	0.642	0.739
	Neural Networks	0.943	0.785	0.533	0.635

TABLE 8. Performance of models using synthetic datasets from CTGAN as training sets.

Synthetic Dataset	Algorithm	Accuracy	Precision	Recall	F1-Score
$\hat{X}_{\text{g},1}$	**Decision Tree**	0.934	0.896	0.830	0.862
	K-NN	0.965	0.973	0.884	0.926
	Logistic Regression	0.918	0.913	0.740	0.817
	Neural Networks	0.953	0.727	0.692	0.709
$\hat{X}_{\text{g},2}$	**Decision Tree**	0.915	0.908	0.758	0.826
	K-NN	0.961	0.940	0.911	0.925
	Logistic Regression	0.953	0.941	0.878	0.908
	Neural Networks	0.936	0.948	0.784	0.858
$\hat{X}_{\text{g},3}$	**Decision Tree**	0.932	0.947	0.931	0.939
	K-NN	0.957	0.963	0.959	0.961
	Logistic Regression	0.846	0.793	0.978	0.876
	Neural Networks	0.935	0.948	0.893	0.920
$\hat{X}_{\text{g},4}$	**Decision Tree**	0.929	0.906	0.805	0.853
	K-NN	0.953	0.933	0.881	0.906
	Logistic Regression	0.941	0.944	0.820	0.878
	Neural Networks	0.938	0.640	0.766	0.697

TABLE 9. Performance of models using SMOTE with either three or five nearest neighbors.

Synthetic Dataset	Algorithm	Accuracy	Precision	Recall	F1-Score
$\hat{X}_{\text{s},1}$ $k = 3$	**Decision Tree**	0.983	0.981	0.985	0.983
	K-NN	0.986	0.981	0.992	0.986
	Logistic Regression	0.955	0.936	0.976	0.956
	Neural Networks	0.921	0.991	0.848	0.914
$\hat{X}_{\text{s},1}$ $k = 5$	**Decision Tree**	0.985	0.982	0.988	0.985
	K-NN	0.984	0.980	0.988	0.984
	Logistic Regression	0.955	0.938	0.974	0.956
	Neural Networks	0.924	0.992	0.852	0.917
$\hat{X}_{\text{s},2}$ $k = 3$	**Decision Tree**	0.965	0.960	0.970	0.965
	K-NN	0.976	0.972	0.979	0.975
	Logistic Regression	0.949	0.930	0.971	0.950
	Neural Networks	0.913	0.995	0.870	0.928
$\hat{X}_{\text{s},2}$ $k = 5$	**Decision Tree**	0.963	0.960	0.966	0.963
	K-NN	0.975	0.976	0.974	0.975
	Logistic Regression	0.951	0.933	0.973	0.953
	Neural Networks	0.908	0.994	0.862	0.923

neighbors to synthesize further instances that are a minor class [54]. One significant difference is that GAN does not require an indication of predetermined classes, whereas SMOTE requires an indication minority classes. As CTGAN is a GAN-based method, the non-convergence probability, or bias of the generator and discriminator should be considered [80]. For SMOTE, the likelihood of overlapping classes, unnecessary noise generation, and high-dimensional datasets should be considered [80]. Thus, if one deals with an imbalanced dataset with an appropriate number of dimensions, SMOTE is a better choice for synthesizing additional samples. However, if one considers a dataset without an apparent class, a dataset that is not tabular data, or wants to gain more quality with less duplication of instances, then GAN is suitable.

In summary, we conclude that the synthetic datasets can train models for tax evasion detection. With more instances, the performance of models becomes higher and this can mitigate the effect of the imbalance in class. This conclusion holds as long as the synthetic datasets represent the relationship of the original data. Labels are not necessary because the result does not differ. Likewise, only independent variables are sufficient to train the models as long as they maintain the original statistical characteristics. In the next section, we conclude our work and discuss possible directions for future work.

V. CONCLUSION AND FUTURE WORK

A. CONCLUSION

Our work proposes a method for obtaining financial data from publicly available sources while respecting privacy laws. This approach contributes to the development of ethical practices for acquiring data from the public. Acquiring tax data is challenging due to its confidentiality and restricted availability. However, studying this area can have a significant impact on society, particularly regarding transparency and equality. Our proposed procedure aims to assist individuals who face challenges related to data availability, enabling them to explore new possibilities in obtaining data.

For our work, we begin with a construction of a dataset based on the collected data to ensure privacy. Subsequently, we studied the risk assessment criteria and assigned labels to the synthetic datasets based on the selected risk assessment criteria. This label is necessary for training machine learning to evaluate whether the synthetic dataset can be used for training instead of a real dataset. Subsequently, the dataset is augmented to obtain a sufficient number of instances. Simultaneously, we considered SMOTE to synthesize a minor class and compared the results with those of CTGAN. From our experiment, we found that as the amount of training data increases, the performances are improved. In terms of synthesis technique, SMOTE provides slightly better results than CTGAN. These synthetic datasets can be used for training machine learning to detect tax evasion if they represented the statistical characteristics of the original dataset. Moreover, applying only independent variables provided sufficient information to train the models. For the criteria to choose between CTGAN and SMOTE for synthesizing a dataset, we concluded that the expected results and further usage are the main criteria. For the appropriate models, we concluded that Decision Tree, K-NN, and Logistic Regression can be trained using the synthetic datasets. Finally, it can be claimed that as the amount of data is limited in today's era, we can use CTGAN as well as SMOTE to generate more data to explore and perform more analysis.

B. FUTURE WORK

After conducting our work, it has become clear that further research in this field is possible. For example, if we consider data related to an asset or cash flow, we can access financial statements, retrieve related data, and apply GAN to obtain sufficient instances. For our proposed dataset, further studies, such as applying other risk assessment criteria, classifying corporation types to see the behavior of different business sectors, or recommending investment decisions, are possible use cases of our collected dataset. Moreover, financial statements from enterprises listed in international stock markets can also be accessed, downloaded, and transformed into tax or other financial-related datasets for further analysis. In addition, if an actual tax return dataset can be obtained, we can evaluate our model further.

For model development, we can do more analysis by controlling the hyperparameters of models to evaluate the stages of models, such as high bias, balance, and high variance. An unsupervised model can be used to explore whether it is capable of clustering the datasets. Furthermore, as a machine learns and is used to detect the current fraud behavior, it can provide more advanced recommendations such as multi-class labeling, suggest definite fraud behavior, or prioritize a potential case for an immediate audit.

Finally, the idea can be applied in fields other than finance as long as public data are available. We hope that our ideas and processes will help others studying tax evasion behavior more, using knowledge from an open source to maximize benefits, increasing instances of data for experiments, and for a better understanding of this field.

REFERENCES

[1] Privacy Internationl (PI). *What is Privacy?* Accessed: Nov. 10, 2022. [Online]. Available: https://privacyinternational.org/explainer/56/what-privacy

[2] H. Dettmar, X. Liu, R. Johnson, and A. Payne, "Knowledge-based data generation," *Knowl.-Based Syst.*, vol. 11, nos. 3–4, pp. 167–177, Nov. 1998, doi: 10.1016/s0950-7051(98)00031-8.

[3] C. Tan, "A model-based approach to generate dynamic synthetic test data," in *Proc. 12th IEEE Conf. Softw. Test., Validation Verification (ICST)*, Xi'an, China, Apr. 2019, pp. 495–497, doi: 10.1109/icst.2019.00063.

[4] The Stock Exchange of Thailand. *Companies/Securities in Focus*. Accessed: Aug. 15, 2022. [Online]. Available: https://www.set.or.th/en/market/get-quote/stock/

[5] Ministry of Finance. *Tax Knowledge*. Accessed: Jun. 23, 2022. [Online]. Available: http://taxclinic.mof.go.th

[6] *Technology Tools to Tackle Tax Evasion and Tax Fraud*, OECD Publishing, Paris, France, 2017.

[7] R. Benkraiem, A. Uyar, M. Kilic, and F. Schneider, "Ethical behavior, auditing strength, and tax evasion: A worldwide perspective," *J. Int. Accounting, Auditing Taxation*, vol. 43, Jun. 2021, Art. no. 100380, doi: 10.1016/j.intaccaudtax.2021.100380.

[8] A. Uyar, K. Nimer, C. Kuzey, M. Shahbaz, and F. Schneider, "Can e-government initiatives alleviate tax evasion? The moderation effect of ICT," *Technol. Forecasting Social Change*, vol. 166, May 2021, Art. no. 120597, doi: 10.1016/j.techfore.2021.120597.

[9] R.-S. Wu, C. S. Ou, H.-Y. Lin, S.-I. Chang, and D. C. Yen, "Using data mining technique to enhance tax evasion detection performance," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 8769–8777, Aug. 2012, doi: 10.1016/j.eswa.2012.01.204.

[10] E. Pappa, R. Sajedi, and E. Vella, "Fiscal consolidation with tax evasion and corruption," *J. Int. Econ.*, vol. 96, pp. S56–S75, Jul. 2015, doi: 10.1016/j.jinteco.2014.12.004.

[11] D. Di Gioacchino and D. Fichera, "Tax evasion and tax morale: A social network analysis," *Eur. J. Political Economy*, vol. 65, Dec. 2020, Art. no. 101922, doi: 10.1016/j.ejpoleco.2020.101922.

[12] M. Savić, J. Atanasijević, D. Jakovetić, and N. Krejić, "Tax evasion risk management using a hybrid unsupervised outlier detection method," *Expert Syst. Appl.*, vol. 193, May 2022, Art. no. 116409, doi: 10.1016/j.eswa.2021.116409.

[13] Tax Justice Network. *The State of Tax Justice 2021*. Accessed: Nov. 10, 2022. [Online]. Available: https://taxjustice.net/reports/the-state-of-tax-justice-2021/

[14] I. Kose, M. Gokturk, and K. Kilic, "An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance," *Appl. Soft Comput.*, vol. 36, pp. 283–299, Nov. 2015, doi: 10.1016/j.asoc.2015.07.018.

[15] V. Nasteski, "An overview of the supervised machine learning methods," *Horizons.B*, vol. 4, pp. 51–62, Dec. 2017, doi: 10.20544/horizons.b.04.1.17.p05.

[16] F. Zhang, B. Shi, B. Dong, Q. Zheng, and X. Ji, "TTED-PU: A transferable tax evasion detection method based on positive and unlabeled learning," in *Proc. IEEE 44th Annu. Comput., Softw., Appl. Conf. (COMPSAC)*, Madrid, Spain, Jul. 2020, pp. 207–216, doi: 10.1109/compsac48688.2020.00036.

[17] A. E. Bouchti, A. Chakroun, H. Abbar, and C. Okar, "Fraud detection in banking using deep reinforcement learning," in *Proc. 7th Int. Conf. Innov. Comput. Technol. (INTECH)*, Luton, U.K., Aug. 2017, pp. 58–63, doi: 10.1109/intech.2017.8102446.

[18] *General Data Protection Regulation GDPR*. Accessed: Nov. 10, 2022. [Online]. Available: https://gdpr-info.eu

[19] *Data Privacy Laws: What You Need to Know in 2022*. Accessed: Nov. 10, 2022. [Online]. Available: https://www.osano.com/articles/data-privacy-laws

[20] Ministry of Digital Economy and Society. *Personal Data Protection Act, B.E. 2562*. Accessed: Nov. 10, 2022. [Online]. Available: https://www.mdes.go.th/law

[21] Nasdaq. *Financials*. Accessed: Nov. 12, 2022. [Online]. Available: https://www.nasdaq.com/ market-activity/quotes/financials

[22] Börse Frankfurt. *Equities*. Accessed: Nov. 12, 2022. [Online]. Available: https://www.boerse-frankfurt.de/equities

[23] T. Peng and A. Telle, "A tool for generating synthetic data," in *Proc. 1st Int. Conf. Data Sci., E-Learn. Inf. Syst.*, Madrid, Spain, Oct. 2018, pp. 1–6, doi: 10.1145/3279996.3280018.

[24] S. A. Assefa, D. Dervovic, M. Mahfouz, R. E. Tillman, P. Reddy, and M. Veloso, "Generating synthetic data in finance: Opportunities, challenges and pitfalls," in *Proc. 1st ACM Int. Conf. AI Finance*, New York, NY, USA, Oct. 2020, pp. 1–8, doi: 10.1145/3383455.3422554.

[25] C. M. Bowen, V. Bryant, L. Burman, S. Khitatrakun, R. McClelland, P. Stallworth, K. Ueyama, and A. R. Williams, "A synthetic supplemental public use file of low-income information return data: Methodology, utility, and privacy implications," in *Proc. Int. Conf. Privacy Stat. Databases*, in Lecture Notes in Computer Science, 2020, pp. 257–270, doi: 10.1007/978-3-030-57521-2_18.

[26] P. Sophaphong, V. Rattanawiboonso, and C. Chanbanjong, "Factors and prevention guideline of tax evasion of listed companies in the stock exchange of Thailand," *J. Humanities Social Sci. Valaya Alongkorn*, vol. 12, no. 3, pp. 47–57, 2017.

[27] W. Didimo, L. Giamminonni, G. Liotta, F. Montecchiani, and D. Pagliuca, "A visual analytics system to support tax evasion discovery," *Decis. Support Syst.*, vol. 110, pp. 71–83, Jun. 2018, doi: 10.1016/j.dss.2018.03.008.

[28] J. Ruan, Z. Yan, B. Dong, Q. Zheng, and B. Qian, "Identifying suspicious groups of affiliated-transaction-based tax evasion in big data," *Inf. Sci.*, vol. 477, pp. 508–532, Mar. 2019, doi: 10.1016/j.ins.2018.11.008.

[29] D. K. Luna, G. K. Palshikar, M. Apte, and A. Bhattacharya, "Finding shell company accounts using anomaly detection," in *Proc. ACM India Joint Int. Conf. Data Sci. Manage. Data*, Goa, India, Jan. 2018, pp. 167–174, doi: 10.1145/3152494.3152519.

[30] J. Kovermann and P. Velte, "The impact of corporate governance on corporate tax avoidance—A literature review," *J. Int. Accounting, Auditing Taxation*, vol. 36, Sep. 2019, Art. no. 100270, doi: 10.1016/j.intaccaudtax.2019.100270.

[31] F. Wang, S. Xu, J. Sun, and C. P. Cullinan, "Corporate tax avoidance: A literature review and research agenda," *J. Econ. Surv.*, vol. 34, no. 4, pp. 793–811, Dec. 2019, doi: 10.1111/joes.12347.

[32] X. Zhu, X. Ao, Z. Qin, Y. Chang, Y. Liu, Q. He, and J. Li, "Intelligent financial fraud detection practices in post-pandemic era," *Innovation*, vol. 2, no. 4, Nov. 2021, Art. no. 100176, doi: 10.1016/j.xinn.2021.100176.

[33] F. Tian, T. Lan, K.-M. Chao, N. Godwin, Q. Zheng, N. Shah, and F. Zhang, "Mining suspicious tax evasion groups in big data," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 10, pp. 2651–2664, Oct. 2016, doi: 10.1109/tkde.2016.2571686.

[34] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: A survey," *Data Mining Knowl. Discovery*, vol. 29, no. 3, pp. 626–688, Jul. 2014, doi: 10.1007/s10618-014-0365-y.

[35] Internal Revenue Service. *History of the Whistleblower Program*. Accessed: Feb. 15, 2022. [Online]. Available: https://www.irs.gov/compliance/history-of-the-whistleblower-program

[36] The Revenue Department. *Tax Evasion Whistleblowing System*. Accessed: Apr. 7, 2023. [Online]. Available: https://interapp61.rd.go.th/taxcomplain2/taxcomplain_landing/home.html

[37] R. Suphati, "Factors affecting on corporate tax refund: A case study of the corporate paying tax exceeding what required for tax refund, revenue department, Chonburi area 2," M.S. thesis, Faculty Manage. Tourism, Burapha Univ., Chonburi, Thailand, 2019.

[38] K. Sektrakul, *Financial Statement Analysis*. Thailand: Thailand Securities Institute (TSI), 2013, p. 44. [Online]. Available: https://classic.set.or.th/dat/vdoArticle/attachFile/AttachFile_1472551305959.pdf

[39] Revenue Department. *Corporate Income Tax*. Accessed: Nov. 12, 2022. [Online]. Available: https://www.rd.go.th/english/6044.html

[40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.

[41] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F. Wang, "Generative adversarial networks: Introduction and outlook," *IEEE/CAA J. Autom. Sinica*, vol. 4, no. 4, pp. 588–598, Sep. 2017, doi: 10.1109/jas.2017.7510583.

[42] A. N. Wu, R. Stouffs, and F. Biljecki, "Generative adversarial networks in the built environment: A comprehensive review of the application of GANs across data types and scales," *Building Environ.*, vol. 223, Sep. 2022, Art. no. 109477, doi: 10.1016/j.buildenv.2022.109477.

[43] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018, doi: 10.1109/msp.2017.2765202.

[44] L. Xu and K. Veeramachaneni, "Synthesizing tabular data using generative adversarial networks," 2018, *arXiv:1811.11264*, doi: 10.48550/arXiv.1811.11264.

[45] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11, doi: 10.48550/arXiv.1907.00503.

[46] B. Brenninkmeijer, A. de Vries, E. Marchiori, and Y. Hille, "On the generation and evaluation of tabular data using GANs," M.S. thesis, Radboud University Nijmegen, The Netherlands, 2019.

[47] S. D. Horn, "Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale," *Int. Biometric Soc.*, vol. 3, no. 1, pp. 237–247, Mar. 1977.

[48] A. A. Mamoon and A. Rahman, "Uncertainty analysis in design rainfall estimation due to limited data length: A case study in Qatar," in *Extreme Hydrology and Climate Variability*. Amsterdam, The Netherlands: Elsevier, 2019, pp. 37–45, doi: 10.1016/B978-0-12-815998-9.00004-X.

[49] F. J. Massey, "The Kolmogorov–Smirnov test for goodness of fit," *J. Amer. Stat. Assoc.*, vol. 46, no. 253, pp. 68–78, Mar. 1951, doi: 10.1080/01621459.1951.10500769.

[50] DataCebo. *Synthetic Data Metrics*. Accessed: Oct. 1, 2022. [Online]. Available: https://docs.sdv.dev/sdmetrics/

[51] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Udupi, India, Sep. 2017, pp. 79–85, doi: 10.1109/icacci.2017.8125820.

[52] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," in *Proc. 11th Int. Conf. Inf. Commun. Syst. (ICICS)*, Irbid, Jordan, Apr. 2020, pp. 243–248, doi: 10.1109/icics49469.2020.239556.

[53] J. Hernandez, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "An empirical study of oversampling and undersampling for instance selection methods on imbalance datasets," in *Proc. Iberoamerican Congr. Pattern Recognit.*, in Lecture Notes in Computer Science, 2013, pp. 262–269, doi: 10.1007/978-3-642-41822-8_33.

[54] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[55] M. Kubat, "Performance evaluation," in *An Introduction to Machine Learning*. Cham, Switzerland: Springer, 2021, ch.12, pp. 233–253, doi: 10.1007/978-3-030-81935-4_12.

[56] F. H. K. dos Santos Tanaka and C. Aranha, "Data augmentation using GANs," 2019, *arXiv:1904.09135*, doi: 10.48550/arXiv.1904.09135.

[57] A. Sethia, R. Patel, and P. Raut, "Data augmentation using generative models for credit card fraud detection," in *Proc. 4th Int. Conf. Comput. Commun. Autom. (ICCCA)*, Greater Noida, India, Dec. 2018, pp. 1–6, doi: 10.1109/ccaa.2018.8777628.

[58] F. Ferreira, N. Lourenco, B. Cabral, and J. P. Fernandes, "When two are better than one: Synthesizing heavily unbalanced data," *IEEE Access*, vol. 9, pp. 150459–150469, 2021, doi: 10.1109/access.2021.3126656.

[59] G. Eilertsen, A. Tsirikoglou, C. Lundström, and J. Unger, "Ensembles of GANs for synthetic training data generation," 2021, *arXiv:2104.11797*, doi: 10.48550/arXiv.2104.11797.

[60] C. Jeong, S. Kim, J. Park, and Y. Choi, "Customs import declaration datasets," 2022, *arXiv:2208.02484*, doi: 10.48550/arXiv.2208.02484.

[61] S. Bourou, A. El Saer, T.-H. Velivassaki, A. Voulkidis, and T. Zahariadis, "A review of tabular data synthesis using GANs on an IDS dataset," *Information*, vol. 12, no. 9, p. 375, Sep. 2021, doi: 10.3390/info12090375.

[62] J. Moon, S. Jung, S. Park, and E. Hwang, "Conditional tabular GAN-based two-stage data generation scheme for short-term load forecasting," *IEEE Access*, vol. 8, pp. 205327–205339, 2020, doi: 10.1109/access.2020.3037063.

[63] P. C. González and J. D. Velásquez, "Characterization and detection of taxpayers with false invoices using data mining techniques," *Expert Syst. Appl.*, vol. 40, no. 5, pp. 1427–1436, Apr. 2013, doi: 10.1016/j.eswa.2012.08.051.

[64] K. Zhang, A. Li, and B. Song, "Fraud detection in tax declaration using ensemble ISGNN," in *Proc. WRI World Congr. Comput. Sci. Inf. Eng.*, Los Angeles, CA, USA, Mar./Apr. 2009, pp. 237–240, doi: 10.1109/csie.2009.73.

[65] C. P. López, M. D. Rodríguez, and S. de Lucas Santos, "Tax fraud detection through neural networks: An application using a sample of personal income taxpayers," *Future Internet*, vol. 11, no. 4, p. 86, Mar. 2019, doi: 10.3390/fi11040086.

[66] L. Zhang, X. Nan, E. Huang, and S. Liu, "Detecting transaction-based tax evasion activities on social media platforms using multi-modal deep neural networks," 2020, *arXiv:2007.13525*, doi: 10.48550/arXiv.2007.13525.

[67] T. Matos, J. A. F. de Macedo, and J. M. Monteiro, "An empirical method for discovering tax fraudsters," in *Proc. 19th Int. Database Eng. Appl. Symp. (IDEAS)*, New York, NY, USA, Jul. 2014, pp. 41–48, doi: 10.1145/2790755.2790759.

[68] D. de Roux, B. Perez, A. Moreno, M. D. P. Villamil, and C. Figueroa, "Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2018, pp. 215–222, doi: 10.1145/3219819.3219878.

[69] J. Vanhoeyveld, D. Martens, and B. Peeters, "Value-added tax fraud detection with scalable anomaly detection techniques," *Appl. Soft Comput.*, vol. 86, Jan. 2020, Art. no. 105895, doi: 10.1016/j.asoc.2019.105895.

[70] R. Wei, B. Dong, Q. Zheng, X. Zhu, J. Ruan, and H. He, "Unsupervised conditional adversarial networks for tax evasion detection," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Los Angeles, CA, USA, Dec. 2019, pp. 1675–1680, doi: 10.1109/bigdata47090.2019.9005656.

[71] Revenue Department. *E-Forms*. Accessed: Aug. 8, 2022. [Online]. Available: https://www.rd.go.th/english/29040.html

[72] sdv-dev GitHub. *CTGAN*. Accessed: Sep. 15, 2022. [Online]. Available: https://github.com/sdv-dev/CTGAN

[73] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Montreal, QC, Canada, Oct. 2016, pp. 399–410, doi: 10.1109/dsaa.2016.49.

[74] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, "YALE," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2006, pp. 935–940, doi: 10.1145/1150402.1150531.

[75] S. Z. Li and A. Jain, "Baseline algorithm," in *Encyclopedia of Biometrics*. Boston, MA, USA: Springer, p. 60, doi: 10.1007/978-0-387-73003-5_538.

[76] A. Alwosheel, S. van Cranenburgh, and C. G. Chorus, "Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis," *J. Choice Model.*, vol. 28, pp. 167–182, Sep. 2018, doi: 10.1016/j.jocm.2018.07.002.

[77] L. E. Burman, A. Engler, S. Khitatrakun, J. R. Nunns, S. Armstrong, J. Iselin, G. MacDonald, and P. Stallworth, "Safely expanding research access to administrative tax data: Creating a synthetic public use file and a validation server," U.S. Internal Revenue Service, Washington, DC, USA, Tech. Rep., 2019.

[78] J. Brownlee, *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*. Vermont, VIC, Australia: Machine Learning Mastery, 2020.

[79] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*, vol. 454. New York, NY, USA: Springer, 2012.

[80] J. Kim and M. Park, "A new body weight lifelog outliers generation method: Reflecting characteristics of body weight data," *Appl. Sci.*, vol. 12, no. 9, p. 4726, May 2022, doi: 10.3390/app12094726.

NARONGCHAI VISITPANYA received the B.E. degree in automotive design and manufacturing engineering from the International School of Engineering (ISE), Chulalongkorn University, Bangkok, Thailand. He is currently pursuing the degree with the Technology of Information System Management Division, Faculty of Engineering, Mahidol University, Nakhon Pathom, Thailand. His research interests include data science and technology related topics.

TAWEESAK SAMANCHUEN received the Ph.D. degree in electrical engineering from the Sirindhorn International Institute of Technology (SIIT), Thammasat University, Thailand. He is currently an Assistant Professor with the Faculty of Engineering, Mahidol University. His current research interests include wireless sensor networks, decision support, analytic hierarchy process (AHP), SIP protocol, embedded systems, computer networks, machine learning, and data science.